Crowdsourcing for Language Resource Development: Criticisms About Amazon Mechanical Turk Overpowering Use

Karen Fort, Gilles Adda, Benoît Sagot, Joseph Mariani, Alain Couillault

Crowdsourcing for Language Resource Development: Criticisms about Amazon Mechanical Turk Overpowering Use

Karën Fort[1], Gilles Adda[2], Benoît Sagot[3], Joseph Mariani[2,4], and Alain Couillault[5]

[1] LORIA & Université de Lorraine, Vandœuvre-lès-Nancy, France
[2] Spoken Language Processing group, LIMSI-CNRS, Orsay, France
[3] Alpage, INRIA Paris–Rocquencourt & Université Paris 7, Rocquencourt, France
[4] IMMI-CNRS, Orsay, France
[5] L3i Laboratory, Université de La Rochelle, France
karen.fort@loria.fr, gilles.adda@limsi.fr, benoit.sagot@inria.fr
joseph.mariani@limsi.fr, alain.couillault@univ-lr.fr

Abstract. This article is a position paper about Amazon Mechanical Turk, the use of which has been steadily growing in language processing in the past few years. According to the mainstream opinion expressed in articles of the domain, this type of on-line working platforms allows to develop quickly all sorts of quality language resources, at a very low price, by people doing that as a hobby. We shall demonstrate here that the situation is far from being that ideal. Our goal here is manifold: 1- to inform researchers, so that they can make their own choices, 2- to develop alternatives with the help of funding agencies and scientific associations, 3- to propose practical and organizational solutions in order to improve language resources development, while limiting the risks of ethical and legal issues without letting go price or quality, 4- to introduce an *Ethics and Big Data Charter* for the documentation of language resources.

Keywords: Amazon Mechanical Turk, language resources, ethics

1 Introduction

Developing annotated corpora, as well as other language resources, involves such high costs that many researchers are looking for alternative, cost-reducing solutions. Among others, crowdsourcing, microworking[6] systems which enable elementary tasks to be performed by a huge number of on-line people, are possible alternatives. Nowadays, Amazon Mechanical Turk (MTurk) is the most popular of these systems, especially in the Speech & Language community. Since its introduction in 2005, there has been a steady growth of MTurk use in building or validating language resources [1].

[6] Microworking refers to the fact that tasks are cut into small pieces and their execution is paid for. Crowdsourcing refers to the fact that the job is outsourced via the web and done by many people (paid or not).

Costs are drastically reduced due to available sparse time of human language experts on-line. But MTurk raises, among others, ethical and quality issues which have been minimized until now, and we will investigate them in this paper. However, because we are aware that the development costs of corpora often stand in the way of language research and technologies, especially for less-resourced languages, we are also sensible of some visible advantages of crowdsourcing. Developing a crowdsourcing system which retains some of the main qualities of MTurk (rapidity, diversity, access to non-expert judgment) while avoiding the ethical and labor laws issues is (theoretically) possible, but this solution will require some delay (in the best case scenario) and the help of our scientific associations (ISCA, ACL, ELRA) and of the national and international funding agencies. Therefore, we will propose existing alternatives aiming at producing high quality resources at a reduced cost, while deliberately keeping ethics above cost savings. In parallel, we created the French *Ethics and Big Data Charter* that will allow funding agencies to select the projects they want to finance according to ethical criteria.

2 MTurk: legends and truth

2.1 MTurk, a hobby for the Turkers?

In order to evaluate the ethics of MTurk, we need to qualify the activity of Turkers while they are participating in MTurk. Is it a voluntary work, as the one in Wikipedia? Looking at the MTurk site or at Turker blogs, where the monetary retribution is a major issue, the answer is clearly no.

Studies in social sciences [2, 3], using surveys submitted within MTurk, give some insight[7] into Turkers' socio-economic facts (country, age...) or the way they use MTurk (number of tasks per week, total income in MTurk ...), and how they qualify their activity. 91% of the Turkers mentioned their desire to make money [5], even if the observed wage is very low;[8] when 60% of the Turkers think that MTurk is a fairly profitable way of spending free time and getting some cash, only 30% mentioned their interest for the tasks, and 20% (5% of the Indian Turkers) say that they are using MTurk to kill time. Finally, 20% (30% of the Indian Turkers) declare that they are using MTurk to make basic ends meet, and about the same proportion that MTurk is their primary source of income.

Looking at the tasks which are performed within MTurk is another way to qualify the Turkers' activity. Innovative kinds of tasks can be found which can be seen as creative hobby activities. However, many tasks correspond to activities which used to be performed by salaried employees, and therefore are working activities; for these tasks, MTurk could be assimilated to off-shoring on the Web

[7] For instance, we learn that Indian Turkers were 5% in 2008, 36% in December 2009 [2], 50% in May 2010 (`http://blog.crowdflower.com/2010/05/amazon-mechanical-turk-survey/`) and have produced over 60% of the activity in MTurk [4].

[8] $1.25/hr according to [6] $1.38/hr according to [7].

to decrease production costs. For years, speech corpora transcription (and translation) tasks were being performed by employees of agencies like LDC or ELDA: these are jobs. The 20% of the most active Turkers who spend more than 15 hours per week in MTurk [8], and produce 80% of the activity, can be called laborers when performing these tasks.

It is difficult to be conclusive about the nature of the Turkers' activity. Many different types of tasks are proposed within MTurk and the Turkers' motivations are heterogeneous. Nevertheless, those 20% of the Turkers for whom MTurk is a primary income, and those Turkers who perform tasks which are actually performed by employees, produce an activity in MTurk corresponding to a real labor.

Qualifying the MTurk activity as labor raises issues about the setup of MTurk. The very low wages (below $2 an hour [6, 3, 7]) are a first point. A further point concerns Amazon's choice of hiding any explicit relationship between Turkers and Requesters, even the basic workplace right of unionization is denied and Turkers have no recourse to any channels for redress against employers' wrongdoing, including the fact that they have no official guarantee of payment for properly performed work. Some regulation between Requesters and Turkers exists through Turkers' blogs or forums[9], or the use of Turkopticon[10], a tool designed to help Turkers reporting bad Requesters. However, all these solutions are unofficial and nothing explicitly protects the Turkers, especially the new ones who are mostly unaware of these tools.

2.2 MTurk drastically reduces costs?

Most articles dealing with MTurk and resource production indicate low costs as the primary motivation. Given the observed salaries (for instance $0.005 to transcribe a 5-second speech segment [9]), the cost may indeed be very low. However, the overall cost is not to be limited to the mere salary: the time needed to develop the interface, and to tackle the spammer problem is not negligible [10]; validation [11] and correction costs [12] to ensure minimal quality are also to be considered. Furthermore, some tasks may become more expensive than expected. This may occur for instance, if the required Turkers' competence is hard to find: to transcribe Korean [9], wages were increased from $5 to $35 per hour.

2.3 MTurk allows for building resources of equivalent quality?

Many technical papers have reported that at least for translation and transcription, the quality is sufficient to train and evaluate statistical translation or transcription systems [10, 13]. However, some of these papers bring to light quality problems.[11]

[9] For instance `mechanicalturk.typepad.com` or `turkers.proboards.com`

[10] `turkopticon.differenceengines.com`

[11] Some of the problems reported, such as the interface problems, are not specific to MTurk, but are generic to many crowdsourcing systems.

Limitations due to the lack of expertise Turkers being non-experts, the requester has to decompose complex tasks into simpler tasks (HITs, Human Intelligence Tasks), to help performing them. By doing so, s/he can be led to make choices that can bias the results. An example of this type of bias is analyzed in [14], where the authors acknowledge the fact that proposing only one sentence per lexical evolution type (amelioration and pejoration) influences the results.

Even more problematic is the fact that the quality produced with MTurk on complex tasks is often not satisfactory. This is for example the case in [15], in which the authors demonstrate that, for their task of word-sense disambiguation, a small number of well-trained annotators produces much better results than a larger group (the number being supposed to counterbalance non-expertise) of Turkers. From this point of view, their results contradict those presented in [16] on a similar, though much simpler, task. The same difficulty arises in [17], in which it is demonstrated that non expert evaluation of summarization systems is "risky", as the Turkers are not able to obtain results comparable to that of experts. More generally, this quality issue can be found in numerous articles in which the authors had to validate Turkers' results using specialists (PhD students in [11]) or use a rather complex post-processing [12]. Finally, the quality of the work from non experts varies considerably [18].

Moreover, there is currently a "snowball" effect going on, that leads to overestimate the resources quality mentioned in articles: some researchers praise MTurk [12], citing research that did use the system, but would not have given usable results without a more or less heavy post-processing [11]. A simplistic conclusion could be that MTurk should only be used for simple tasks, however, besides the fact that MTurk itself induces important limitations (see next section), it is interesting to notice that, for some simple tasks, Natural Language Processing tools already provide better results than the Turkers [19].

Limitations due to MTurk itself In [18], the authors note that the limits of the user interface constitute the "first and most important drawback of MTurk". The authors also regret that it is impossible to be 100% sure that the Turkers participating in the task are real native English speakers. If pre-tests can be designed to address, at least partly, this issue, they represent an added cost and it will still be very easy to cheat [10]. Of course, one can always organize various protections [10], but here again, this requires time and therefore represents an additional cost that only few requesters are ready to pay for.[12] For example, in [12], the authors identified spammers but did not succeed in eliminating them.

Finally, the impact of task payment should not be neglected, as it induces as logical behavior to place the number of performed tasks above quality, regardless of payment. In [20] the authors thus reached the conclusion that an hourly payment was better (with some verification and time justification procedures).

[12] Interestingly, it seems that MTurk recently decided to no longer accept the non-US Turkers, for quality and fraud reasons: `http://turkrequesters.blogspot.fr/2013/01/the-reasons-why-amazon-mechanical-turk.html`.

3 Existing or suggested alternatives

MTurk is not the only way to achieve fast development of high quality resources at a low cost. First, and despite the lack of systematic studies, existing automatic tools seem to perform as well as (non-expert) Turkers, if not better, on certain tasks [19]. Second, the cost of tasks like manual annotation can be drastically reduced using the appropriate techniques. Third, exploiting as much as possible existing resources can be an inexpensive alternative to MTurk. Finally, MTurk is not the only crowdsourcing and microworking platform.

3.1 Unsupervised and semi-supervised techniques for low-cost language resource development

Unsupervised machine learning techniques have been studied in the Speech & Language community for quite a long time, for numerous and sometimes complex tasks, including tokenization, POS tagging [21], parsing [22] or document classification. Although such techniques produce results that are below state-of-the-art supervised or symbolic techniques, which both require resources that are costly to develop, it is unclear whether they produce results that are below what can be expected from MTurk, especially for complex tasks such as parsing. Moreover, unsupervised techniques can be improved at a reasonable cost by optimizing the construction and use of a limited amount of additional information (annotations, external resources). This constitutes the **semi-supervised learning** paradigm [23]. Such approaches for developing language resources rely on two (complementary) principles:

- Training models on a limited amount of annotated data and using the result for producing more annotation. For example, using one model, one can select within the automatically annotated data those that have a high confidence level, and consider that as additional training data (*self-training*, [24]). Using two different models allows to rely on the high-confidence annotations of one model for augmenting the training corpus for the other, thus decreasing systematic biases (*co-training*, [25]). If one accepts to produce a limited amount of manual annotations not only in advance but also while developing the tools, one can request the manual annotation of carefully chosen data, i.e., data for which knowing the expected output of the system improves as much as possible the accuracy of the system (*active learning* [26]).
- Using data containing annotations that are less informative, complete and/or disambiguated than the target annotations, like a morphological lexicon (i.e., an ambiguous POS-annotation) for POS tagging [27], a morphological description for morphological lexicon induction [28] or a partly bracketed corpus for full parsers [29].

3.2 Optimizing the cost of manual annotation: pre-annotation and dedicated interfaces

When using approaches that rely on expert annotation, this annotation can be sped up and sometimes even improved by automatic annotation tools used as

pre-annotators. For instance, [30] have shown that for POS tagging, a low-quality and non-costly pre-annotation tool can drastically improve manual annotation speed; 50 manually POS-annotated sentences are enough for training a pre-annotation tool that reduces manual work as much as a state-of-the-art POS tagger, allowing to developing a 10,000-sentence standard-size corpus in ∼100 hours of expert work. On the other hand, on such a task, one could question the ability of anonymous Turkers to correctly follow detailed and complex annotation guidelines.

Obviously, the above-mentioned remarks by [18] about the limitations of MTurk interfaces apply more generally. Past projects aiming at developing syntactically and semantically annotated corpora have shown that both the speed and quality of the annotation is strongly influenced by the annotation interface itself [31]. This provides another source of improvements for annotation efficiency and quality. Put together, it might well be the case that even costly expert work can be used in optimized ways that lead to high-quality resources at a reasonable cost, even compared with that of MTurk.

3.3 Reusing existing resources

Even less costly is the **use of existing data** for creating new language resources. An example is the named-entity recognition (NER) task. MTurk has been used for developing NER tools, in particular for specific domains such as medical corpora [32], twitter [33] or e-mails [34]. However, converting Wikipedia into a large-scale named-entity-annotated resource leads to building high-quality NER tools [35], including when evaluated on other types of corpora [36]. Apart from Wikipedia (and the related DBpedia), other wiki projects (e.g., wiktionaries) and freely-available resources (lexicons, corpora) are valuable sources of information.

3.4 Collaborative or crowdsourced development beyond MTurk

All these alternatives require a fair amount of expert work. Other approaches do exist that reduce this requirement to a low level, and in particular collaborative and game-based techniques, as well as crowdsourcing platforms other than MTurk, which try to avoid at least in part its pitfalls.

Collaborative approaches for language resource development rely on the strategy set up by the Wikipedia and other Wikimedia projects, as well as other wikis such as semantic wikis (Freebase, OntoWiki...). Anyone can contribute linguistic information (annotation, lexical data...), but usually contributors are motivated because they are to some extent experts themselves. The quality control is usually done mutually by contributors themselves, sometimes by means of on-line discussions, often leading to high quality results. One of the first collaborative platforms for language resource development was the semantic annotation tool Serengeti [37], currently used within the AnaWiki project.[13]

[13] `http://www.anawiki.org`

However, such approaches remain more suitable for developing medium-scale high-quality resources. For the fast development of large-scale resources, another strategy is to attract a large number of non-experts thanks to online games, that fall in the family of so-called **games with a purpose (GWAP)**. This idea was initiated by the ESP on-line game [38] for image tagging. Its success led researchers to develop such games for various tasks, including language-related ones. A well-known example is *PhraseDetective* [39] for annotating anaphoric links, a reputedly complex task, which lead the authors to include a training step before allowing players to actually provide new annotations. However, the boundary between GWAPs and crowdsourcing is not clear-cut. It is not the case that MTurk remunerates a work whereas other approaches are purely "for fun". Indeed, even contributing to Wikipedia is a job, though a voluntary unpaid job. GWAP and MTurk cannot be distinguished either by the fact that MTurk gives a remuneration, as some GWAPs do propose non-monetary rewards (e.g. Amazon vouchers for PhraseDetective). Finally, collaborative and GWAP-based techniques are not the only "ethical alternatives", since ethical crowdsourcing platforms do exist.

For gathering language data, in particular for less-resourced languages, **crowdsourcing platforms apart from MTurk** seem to be particularly appropriate, as shown for example by speech corpus acquisition experiments using dedicated applications run on mobile phones [40]. An example of an ethical crowdsourcing platform is Samasource, an NGO that allows really poor people to be properly trained and paid for specific tasks (e.g. translating SMS in Creole after the earthquake in Haiti for helping victims and international rescuers to communicate).[14]

4 Towards traceability: the *Ethics and Big Data Charter*

To adopt an ethical behavior in developing, funding, using or promoting language resources is first and above all a matter of choice: for the provider, deciding which approach to adopt – crowdsourcing or not –, or which platform to request on, or the level of remuneration of the workers; for the funding agency, choosing which project to fund; for users, choosing which resource to use or acquire. These choices have to be learned ones. We designed the *Ethics and Big Data Charter*[41] in collaboration with representatives of interest groups, private companies and academic organizations, including the French CNRS[15], ATALA[16], AFCP[17] and APROGED [18]. The purpose of this charter is to provide resources developers with a framework to document their resources and ensure their traceability and transparency.

[14] `http://www.samasource.org/haiti/`

[15] Centre National de la Recherche Scientifique/National agency for scientific research

[16] Association pour le Traitement Automatique des Langues/Natural Language Processing Association `http://www.atala.org`

[17] Association Française de Communication Parlée/French spoken communication association, `http://www.afcp-parole.org`

[18] Association de la Maîtrise et de la Valorisation des contenus/Association for mastering and empowering content, `http://www.aproged.org`

4.1 Why Big Data?

In the process of writing the Charter, it soon appeared that the issues raised for language resources apply to a larger range of data sets, which can be described as Big Data. Indeed, Big Data are characterized not only by their volume, but also by the complexity of the data, which is in no doubt the case even for small sets of language resources. Reversely, the reflexions conducted for language resources can be generalized to and benefit to Big Data sets.

4.2 Contents of the Charter

The *Ethics and Big Data Charter* is provided as a form to be filled in by the dataset provider. It is split into three major sections: *traceability*, *intellectual property* and *specific legislation*, preceded by a short identification section containing the names of the resource, the contact and responsible persons and a short description of the data set.

Traceability *Traceability* is key to our purpose of putting forward ethical issues. The traceability part of the charter allows to precise the relationship between the resource provider and the workers involved in developing the resource, including legal bounding, workers skills, selection criteria.
Specific focus is put on personal data, i.e. data, like voice or video recording, which can provide a means to identify a person directly or indirectly. The Charter requires to precise if and how the data is de-identified, and if and how the individuals were informed of the purpose of the data collection.
Quality assurance is another major aspect of traceability addressed by the charter, as it requires to document the quality assurance strategy, so that the user of the data set is fully informed on the level of quality s/he can expect: what QA procedure the data were passed through? what portion of the data has been evaluated? What are the actual metrics used and their values?

License and copyright Thanks to a great deal of effort accomplished in the definition of – mainly open source – license schemes, it has become common practice to attach a license to a data set. The License and Copyright section of the Charter goes beyond this and puts the focus on questions which may be disregarded, like ensuring that the legal or moral copyrights of the persons who worked on compiling, enriching or transforming the data are respected. As an example, we saw to it that all the writers of the *Ethics and Big Data Charter* are mentioned in the license citation. Also, the Charter reminds data collectors and distributors that they should check whether they comply with any third party data license they may use.

Specific legal requirements A third section of the *Ethics and Big Data Charter* deals with legal requirements that may arise from certain properties of the data set. For example, a country may have issued specifics laws regarding the

storing, use and/or dissemination of personal data. The Charter serves as a reminder for checking if such requirements exist.

4.3 Availability

The *Ethics and Big Data Charter* is available on-line.[19] The website is currently in French, and an English translation of the *Ethics and Big Data Charter* is available.[20]
Examples of charters are also provided, including one for a corpus of e-mail messages, and one for a medical dataset. Both corpus raise privacy issues that the *Ethics and Big Data Charter* allows to deal with.

5 Conclusion and perspectives

We have tried to demonstrate here that MTurk is no panacea and that other solutions exist allowing to reduce the development costs of high-quality language resources, while respecting those working on the resources and their skills.

We would like, as a conclusion, to go beyond the present facts and insist on the longer term consequences of this trend. Under the pressure of this type of low-cost systems, funding agencies could become more reluctant to finance language resources development projects at "normal" costs. The MTurk cost would then become a *de facto* standard and we would have no other choice as for the development method.

We saw, in section 3.3, that a microworking system can generate paid tasks while preserving ethics. This can even represent a chance for people who cannot participate in the usual labor market, due to their remoteness, their handicap, etc., but it requires a strict legal framework to ensure that the system does not violate their rights as workers. This is why we propose that the concerned associations, like the ACL[21] for natural language processing, the ISCA[22] for speech and the ELRA[23] for Language Resources take care of this problem and push to the development and dissemination of the needed tools to better qualify the quality and ethics of the language resources, such as the *Ethics and Big Data Charter*. For that purpose, we already engaged with funding agencies at the French level, some of which have adopted the charter as part of their projects selection process. This effort would need to be extended to international organizations.

Acknowledgments. This work was partly realized as part of the Quæro Programme, funded by OSEO, French State agency for innovation, as well as part of the French ANR project EDylex (ANR-09-CORD-008) and of the Network of

[19] http://wiki.ethique-big-data.org
[20] http://wiki.ethique-big-data.org/chartes/charteethiqueenV2.pdf
[21] http://www.aclweb.org/
[22] http://www.isca-speech.org/
[23] http://www.elra.info/

Excellence "Multilingual Europe Technology Alliance (META-NET)", co-funded by the 7th Framework Programme of the European Commission through the contract T4ME (grant agreement no.: 249119).
We would like to thank the authors[24] of the *Ethics and Big Data Charter* for their dedicated time and effort.

References

1. Fort, K., Adda, G., Cohen, K.B.: Amazon mechanical turk: Gold mine or coal mine? Computational Linguistics (editorial) **37**(2) (2011)
2. Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers?: shifting demographics in mechanical turk. In: Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems. CHI EA '10, New York, NY, USA, ACM (2010)
3. Ipeirotis, P.: Demographics of mechanical turk. CeDER Working Papers, http://hdl.handle.net/2451/29585 (March 2010) CeDER-10-01.
4. Biewald, L.: Better crowdsourcing through automated methods for quality control. SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (January 2010)
5. Silberman, M.S., Ross, J., Irani, L., Tomlinson, B.: Sellers' problems in human computation markets. In: Proceedings of the ACM SIGKDD Workshop on Human Computation. HCOMP '10 (2010) 18–21
6. Ross, J., Zaldivar, A., Irani, L., Tomlinson, B.: Who are the turkers? worker demographics in amazon mechanical turk. Social Code Report 2009-01, http://www.ics.uci.edu/ jwross/pubs/SocialCode-2009-01.pdf (2009)
7. Chilton, L.B., Horton, J.J., Miller, R.C., Azenkot, S.: Task search in a human computation market. In: Proceedings of the ACM SIGKDD Workshop on Human Computation. HCOMP '10 (2010) 1–9
8. Adda, G., Mariani, J.: Language resources and amazon mechanical turk: legal, ethical and other issues. In: LISLR 2010, "Legal Issues for Sharing Language Resources workshop", LREC 2010, Valletta, Malta (May 2010)
9. Novotney, S., Callison-Burch, C.: Cheap, fast and good enough: automatic speech recognition with non-expert transcription. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT '10, Los Angeles, California, USA (2010) 207–215
10. Callison-Burch, C., Dredze, M.: Creating speech and language data with amazon's mechanical turk. In: CSLDAMT '10: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, California, USA (2010)
11. Kaisser, M., Lowe, J.B.: Creating a research collection of question answer sentence pairs with amazon's mechanical turk. In: Proceedings of the International Language Resources and Evaluation Conference (LREC), Marrakech, Morocco (2008)
12. Xu, F., Klakow, D.: Paragraph acquisition and selection for list question using amazon's mechanical turk. In: Proceedings of the International Language Resources and Evaluation Conference (LREC), Valletta, Malta (May 2010) 2340–2345
13. Marge, M., Banerjee, S., Rudnicky, A.I.: Using the amazon mechanical turk for transcription of spoken language. In: IEEE International Conference on Acoustics

[24] `http://wiki.ethique-big-data.org/index.php?title=Ethique_Big_Data:Accueil#Les_auteurs`

Speech and Signal Processing (ICASSP), Dallas, USA (14-19 March 2010) 5270–5273
14. Cook, P., Stevenson, S.: Automatically identifying changes in the semantic orientation of words. In: Proceedings of the International Language Resources and Evaluation Conference (LREC), Valletta, Malta (May 2010)
15. Bhardwaj, V., Passonneau, R., Salleb-Aouissi, A., Ide, N.: Anveshan: A tool for analysis of multiple annotators' labeling behavior. In: Proceedings of The fourth linguistic annotation workshop (LAW IV), Uppsala, Sweden (2010)
16. Snow, R., O'Connor, B., Jurafsky, D., Ng., A.Y.: Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of EMNLP 2008. (2008) 254–263
17. Gillick, D., Liu, Y.: Non-expert evaluation of summarization systems is risky. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. CSLDAMT '10, Los Angeles, California, USA (2010)
18. Tratz, S., Hovy, E.: A taxonomy, dataset, and classifier for automatic noun compound interpretation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden (July 2010) 678–687
19. Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., Marin, D., Simons, H.: Towards building a high-qualityworkforce with mechanical turk. In: Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS). (December 2010)
20. Kochhar, S., Mazzocchi, S., Paritosh, P.: The anatomy of a large-scale human computation engine. In: Proceedings of Human Computation Workshop at the 16th ACM SIKDD Conference on Knowledge Discovery and Data Mining, KDD 2010, Washington D.C. (2010)
21. Goldwater, S., Griffiths, T.: A fully bayesian approach to unsupervised part-of-speech tagging. In: Proceedings of ACL, Prague, Czech Republic (2007)
22. Hänig, C.: Improvements in unsupervised co-occurrence based parsing. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. CoNLL '10, Uppsala, Sweden (2010) 1–8
23. Abney, S.: Semisupervised Learning for Computational Linguistics. 1ère edn. Chapman & Hall/CRC (2007)
24. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, USA (1995) 189–196
25. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers (1998)
26. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. In Tesauro, G., Touretzky, D., Leen, T., eds.: Advances in Neural Information Processing Systems. Volume 7., The MIT Press (1995) 705–712
27. Smith, N., Eisner, J.: Contrastive estimation: Training log-linear models on unlabeled data. In: Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, USA (2005) 354–362
28. Sagot, B.: Automatic acquisition of a Slovak lexicon from a raw corpus. In: Lecture Notes in Artificial Intelligence 3658 ((c) Springer-Verlag), Proceedings of TSD'05, Karlovy Vary, Czech Republic (2005) 156–163
29. Watson, R., Briscoe, T., Carroll, J.: Semi-supervised training of a statistical parser from unlabeled partially-bracketed data. In: Proceedings of the 10th International Conference on Parsing Technologies. IWPT '07, Prague, Czech Republic (2007)

30. Fort, K., Sagot, B.: Influence of Pre-annotation on POS-tagged Corpus Development. In: Proc. of the Fourth ACL Linguistic Annotation Workshop, Uppsala, Sweden (2010)
31. Erk, K., Kowalski, A., Pado, S.: The salsa annotation tool. In Duchier, D., Kruijff, G.J.M., eds.: Proceedings of the Workshop on Prospects and Advances in the Syntax/Semantics Interface, Nancy, France (2003)
32. Yetisgen-Yildiz, M., Solti, I., Xia, F., Halgrim, S.R.: Preliminary experience with amazon's mechanical turk for annotating medical named entities. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. CSLDAMT '10, Los Angeles, California, USA (2010) 180–183
33. Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M.: Annotating named entities in twitter data with crowdsourcing. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. CSLDAMT '10, Los Angeles, California, USA (2010)
34. Lawson, N., Eustice, K., Perkowitz, M., Yetisgen-Yildiz, M.: Annotating large email datasets for named entity recognition with mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. CSLDAMT '10, Los Angeles, California, USA (2010) 71–79
35. Nothman, J., Curran, J.R., Murphy, T.: Transforming Wikipedia into Named Entity Training Data. In: Proceedings of the Australian Language Technology Workshop. (2008)
36. Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., Curran, J.R.: Named entity recognition in wikipedia. In: People's Web '09: Proceedings of the 2009 Workshop on The People's Web Meets NLP, Suntec, Singapore (2009) 10–18
37. Stürenberg, M., Goecke, D., Die-wald, N., Cramer, I., Mehler, A.: Web-based annotation of anaphoric relations and lexical chains. In: ACL Workshop on Linguistic Annotation Workshop (LAW), Prague, Czech Republic (2007)
38. von Ahn, L.: Games with a purpose. IEEE Computer Magazine (2006) 96–98
39. Chamberlain, J., Poesio, M., Kruschwitz, U.: Phrase Detectives: a Web-based Collaborative Annotation Game. In: Proceedings of the International Conference on Semantic Systems (I-Semantics'08), Graz, Austria (2008)
40. Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P., LeBeau, M.: Building transcribed speech corpora quickly and cheaply for many languages. In: Proceedings of Interspeech, Makuhari, Chiba, Japan (September 2010) 1914–1917
41. Couillault, A., Fort, K.: Charte Éthique et Big Data : parce que mon corpus le vaut bien ! In: Linguistique, Langues et Parole : Statuts, Usages et Mésusages, Strasburg, France (July 2013) 4 pages.